

# Voice UI における時間要素の制御

神場知成<sup>†1</sup> 中林寿文<sup>†2</sup>

**概要:** スマートスピーカーが急速に普及するなど、Voice UI (VUI) がさまざまな場面で利用されるようになってきているが、これはおもに音声認識・合成技術とインターネット技術の進化とが結びついたもので、ヒューマンインタフェースの側面から見ると、さまざまな改良の余地がある。本論文では、VUI システムの発話の間(ま)の取り方、発話スピードなど時間要素の制御が、人間にとっての心地よさに与える影響に着目した実験を行うことでその重要性を示すとともに、その制御をするための指針について述べる。

## 1. はじめに

Amazon Echo などのスマートスピーカーに代表される、VUI を利用するシステムが急速に普及し始めているが、現時点での VUI は、一般の利用者が期待する「あたかも人間と対話するように、自由に声で対話できるシステム」には遠く、そのギャップを埋めるためにさまざまな研究が行われている。日常生活の中で人間が VUI に話しかける様子をエスノグラフィ的に観察・分析する研究[8]や、人工知能の高度化により自然言語による知的で複雑な会話を可能にするための研究[15]などである。しかし、ヒューマンインタフェースの観点から見れば、これまで扱われていない、VUI システムに特有の課題が多々残されていると考えている。その一つとして本論文で提示するのは、システムの発話タイミングや発話速度などの時間要素である。

たとえば、朝、子供が家を出るときに親に「今日、傘を持って行った方がいいかな？」と聞いた場合を考えてみよう。親が即座に「その方がいいよ」と答えた場合と、3 秒経ってから「その方がいいよ」と答えた場合とで、子供は前者の場合に、より強く傘の必要性を感じるであろう。同じ発話であっても、どのようなタイミングで言うか、どのようなスピードで言うか、などの時間要素は非常に重要であるが、従来の VUI システムを見ても、このような考慮をしたシステム設計はほとんど行われていない。

その原因の一つは、現在の商用システムにおいて、これらの時間要素は、端末とネットワーク間での通信時間、システム内部の処理時間などの要素で決まり、技術的に簡単に制御することができないと考えられているからであろう。

本論文は、VUI システムからの発話タイミングを数百ミリ秒単位で制御したときに、人間の受け取り方どのように影響するかを調べるものである。

## 2. 既存研究

音声だけでなくテキストも含めた自然言語対話の研究は古く、チューリングテストに遡る[11]。これは、相手が誰であるかわからない状態で、人間がその相手と自然言語でやりとりをし、機械であるか人間であるかを見分けるこ

とができなければ、そのマシンは知的だと呼べる、というものである。これについては、2014 年に、あまり知的とは言えない対話プログラムが史上初めて合格したこともあり[1]、現在では以前ほど強く支持されていないが、少なくとも、人工知能の一つの基準として自然言語による対話能力が重視され、その研究の進展に貢献したことは確かである。

さらに 1960 年代末、Winograd が SHRDLU を開発したことも、自然言語対話におけるエポックメイキングな出来事の一つである[13]。SHRDLU は、積み木の部屋という限定された局面を設定することにより、談話の問題、指示語、曖昧性の問題などを扱い、連続する会話によって、積み木に関する内部モデルを適切に変更することができる。

SHRDLU の後、自然言語対話技術は進歩するというよりはむしろ、困難さが明らかになりつつあったが、音声認識、音声合成技術の進歩がインターネットと組み合わせることでスマートスピーカーは近年、急速に実用化が進んだ。レポートによれば[3]、2018 年の第一四半期にスマートスピーカーは世界で 900 万台売れ、前年比 210% の伸びを示し、2018 年 8 月の別のレポートでは[2]、米国で利用されているスマートスピーカーは 5000 万台と報告されている。

技術的には発展途上であるとは言え、これだけ利用されるようになると当然、人間の生活への影響は大きい。ACM の CHI 2018 では Voice UI に関するワークショップが行われてさまざまな観点から議論がされている[9]。その中の 1 つとして、Tsai 等がさまざまな人が VUI をどのような目的に利用したいかを調べた中で[10]、「寂しいときや退屈なときに話し相手になってもらいたい」という人たちが一定以上いることは興味深い。筆者自身も、大学 1 年生 140 人に対して「音声 UI を何に使いたいか」というアンケートをとった時に「雑談、気軽なおしゃべり」は最も多い答の 1 つであった[5]。

これは、VUI では単に必要な情報を伝達することや正確な回答を返すことだけが重要なわけではないということも意味している。発話によって受信者が受け取る印象や、その感情への影響も極めて重要となってくる可能性がある。本研究では、そのような、VUI の発話が人間に与える印象が、主として時間的な要素により、どのような影響を受けるかについて述べる。

<sup>†1</sup> 東洋大学 情報連携学部 (INIAD)

<sup>†2</sup> (株) サイバード

### 3. 実験用システム

ユーザと VUI システムとの会話において、ユーザの発話に対する VUI のレスポンスタイミングがユーザの印象に与える影響を調べるため、次のような実験を計画した。

- 人間と VUI システムとの間で会話をしてもらい、システム発話のパラメータ設定の相違により人間が受ける印象の違いを調べる。
- 情報を得るためだけの会話と、心情的なやりとりの会話との2つのシーンを想定する。
- それぞれのシーンにおいて、VUI システムからのレスポンスは、設定パラメータである発話タイミングと発話スピードの組み合わせをさまざまに変える。さらにピッチの影響も知るため、一部変更する。
- 利用者には、それぞれの設定パラメータでのレスポンスから受け取る印象を、心地よい (comfortable) から不快 (uncomfortable) まで4段階で選択してもらう。

これにより、VUI システムがどのようなタイミング、スピードでレスポンスをするときにユーザが心地よい、または不快と感じるかを調べる。実装はブラウザ上に HTML/CSS および JavaScript で行い、実験は PC の入力出力を用いて行うものとする。具体的な設定を、次に示す。

シーン：

「コンビニに、店員の代わりに VUI がある」「自宅に、会話相手として VUI がある」という2つのシーンを設定する。前者を「コンビニ」、後者を「家」のシーンと呼ぶ。実験として行った対話のセリフは、図1に示すもので、U は User、S は System を示す。

U: 炭酸水はありますか？  
S: はい、右奥の棚のところにあります。  
U: すいません、どこの棚ですか？  
S: その列の一番奥のところです。  
U: それでは、3本ください。  
S: はい、900円になります。

(a) 「コンビニ」のシーン

U: 今日、会社で嫌なことがあって  
S: どうしたの？  
U: ミスをして、上司におこられたんだ。  
S: それは大変だったね。  
U: 昨日も怒られたんだよ。  
S: 気にしない方がいいよ。

(b) 「家」のシーン

図1 VUI システムとの会話シナリオ

### パラメータ

VUI システムは、次のように3つのパラメータを変化させる。主に変更するのはインターバルと発話スピードであり、ピッチは、一部のみ変更して実験を行う。

- インターバル：ユーザ発話の一部（特定の単語またはフレーズ）をシステムが認識してから、システムが発話を始めるまでの時間を制御する。1200ms の相違があるもの3パターン（つまり、最大で2400secの差がある）。ユーザ発話のどの部分を認識したらシステム内部で発話準備を開始するかを、図1の文に下線を引いて示している。
- 発話スピード：遅いものから速いものまで3段階
- ピッチ：低いものから高いものまで3段階。

ユーザ発話の認識とシステムによる発話は、W3C の Speech API Community Group 発表による Web Speech API Specification [7] に基づいて Chrome 上に実装されている API (webkitSpeechRecognition と SpeechSynthesisUtterance) を JavaScript から呼び出すことで実装している。

上記で発話スピードとピッチを設定するために、具体的には、JavaScript から SpeechSynthesisUtterance で生成したオブジェクトに対し、次のような値を設定している。

- スピード：1, 2, 4
- ピッチ：0, 0.9, 1.8

API 仕様の上では、スピードは0（もっとも遅い）から10（もっとも速い）までの整数が設定可能だが、日常会話のスピードとして0は遅すぎ、4以上は速すぎて不自然と考えたため、最高レベルを4に設定した。また、ピッチに関しては0（もっとも低い）から2（もっとも高い）までの小数が設定可能だが、2は通常の声としては非常にピッチが高いものとなるため、基本的な実験は0.9に設定して行い、一部の実験で、0および1.8と設定した。



図2 実験システムの画面インタフェース

図2が実験システムの画面インタフェースである。実験者は被験者の横で被験者に操作指示を与え、被験者はそれ

に基づいて作業を行う。

被験者は図2の画面左上部でシーン選択を行う。そこには、コンビニおよび家というそれぞれの設定名が書いてあり、設定名を選択すると、それがどのようなシーンなのかを想像させるための写真が表示される。中央部には、それぞれのシーンの中で被験者が VUI システムに対して言うべき発話内容が、順に書かれている。その時の VUI システムからのレスポンスは、「……」としてマスクされており、画面には表示していない。

画面右側には、システムの発話パラメータとして異なる設定がされた、5つのセット(A~E)が縦に並んでいる。それぞれのセットの中に、さらに3種類のパラメータ設定がある。たとえば、セットAに属する3つの設定はすべて発話スピードとピッチが同じ値に設定されているが、インターバルは異なる3つの値に設定されている。セットBに属する3つの設定は、セットAとはそれぞれ発話スピードだけが異なる。実験者は、AからEまでセットを順番に指示し、被験者は、それぞれのセット中に属するパラメータ設定を順番に試す。1つの設定を選んだうえで、画面に書かれているように発話をし、それに対する VUI システムのレスポンスを聞いて次の発話を行う。それが終了したら右下のエリアで、そのパラメータ設定に対する評価を4(心地よい)から1(不快)までの4段階で選択し、次の設定を試すためにパラメータ設定を変更する。1つのセットの中で、一度パラメータ設定を評価した後、別のパラメータ設定を聞いたうえで、すでに行ったパラメータ設定を再度試し、評価を変更することも許されている。この作業をA~Eの5つのセットに対して行う。1つのシーンに対してこれが終了すると、もう1つのシーンに対して行う。

つまり被験者は、5セット(1セットには、3種類のパラメータ設定がある)×2シーンで合計30個のパラメータ設定に対し、心地よさを4段階で選択することになる。

図3は、システム全体のアーキテクチャと、発話インターバルの制御方法の説明である。Chrome ブラウザは、通信内容の詳細は公開されていないものの、音声認識をするにあたりサーバとのコミュニケーションを行う[6]。利用者がパラメータ設定を選択した上で発話をする、システムは音声認識をして、それに対するレスポンスを返す。レスポンスのタイミング、スピードやピッチは、その時に選択されているパラメータの値に従う。

以下で、図3を用いて、発話インターバルに対してさらに詳しく説明する。本実験システムは、被験者の発話のすべてを認識してからレスポンスを返すわけではない。利用者が発話を始めてから、特定のフレーズや単語を認識したら、その時点から、あらかじめ設定した時間が経過したら発話を始める。これは、たとえば、「炭酸水はありますか?」という発話において、冒頭にある「炭酸水」という言葉を認識しただけで、VUI システムはレスポンスの発話準備を開始するように設定している。たとえば上記の文では「炭酸水」という単語を認識した時からどれだけインターバルを入れて発話コマンド(Javascriptにおける`speechSynthesis.speak(text)`)を発行するかを制御している。本システムの中では、該当する単語を認識してから実際にレスポンスの指示をするまでに400ミリ秒の遅延を入れている。理由は、これにより利用者の発話終了とほぼ同時にレスポンスが開始するような発話文の設計になっているからであるが、これはあくまでアドホックな数値なので、以下でシステムの遅延時間を説明する際には、この400ミリ秒は除いて示す。

インターバルは400ミリ秒単位で制御しており、0.0秒、1.2秒、2.4秒の3種類で実験を行った。たとえば、被験者の「炭酸水はありますか?」という発話に対して、インターバルのパラメータを最小の0.0秒に設定した場合と最大の2.4秒に設定した場合を比較すれば、後者は前者よりも2.4秒遅れて発話することになる。

なお、ここで述べたように、文全体を認識してからシステムが発話を開始するのではなく、冒頭の単語を認識しただけで発話するというのは実験システムとしてのアドホックな方法であるが、実際の人間の会話においても用いられている合理的方法であると考えている。人間であれば、相手が話をしている途中であっても、一部の発話だけから全体の意味をほぼ正しく推定し、それに対する答を言い始めることがよくある。ここで示した実験のように、コンビニのレジで顧客が「炭酸水はありますか?」という発話をするならば、「炭酸水」と言った瞬間に、店員は実際にそれが置いてある棚の場所を伝えるべきだと理解する可能性は非常に高く、それを実際に口に出すタイミングを、意図的に制御していると考えられるからである。

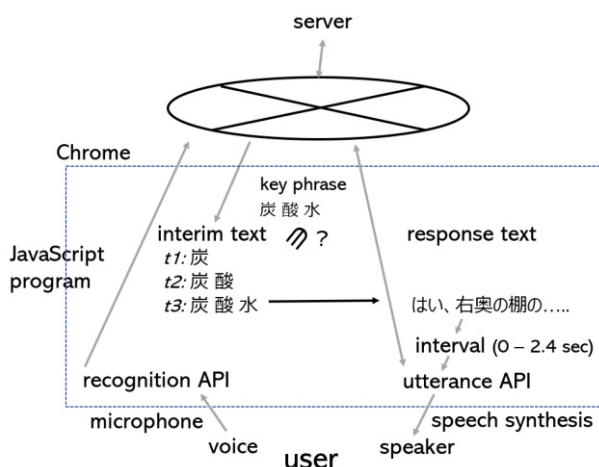


図3 実験システムのアーキテクチャ

```

speechstart: 0
temp: パン: 886
temp: 炭酸: 899
temp: 炭酸水: 1163
reply_to 炭酸水: 1164
temp: 炭酸水を: 1275
temp: 炭酸水は: 1285
temp: 炭酸水はあり: 1471
temp: 炭酸水はありま: 1573
temp: 炭酸水はあります: 1581
temp: 炭酸水はありますか: 1780

```

図4 システムが出力するログの例

システムは、図4に示すようなログを出力する。この例では、最初の `speechstart` というところで、人間の発話開始を検知すると同時にタイマーを0にセットし、各行の最後にある数字が、そこから経過した時間をミリ秒で示している。行頭の `temp` というのは認識途中（認識結果が未確定）であることを示し、その時の暫定的な認識結果を表示している。具体的には、`SpeechRecognitionEvent` における `results attribute` の `array` の要素を参照している。2行目では「炭酸水」の「炭」という部分をシステムが、少し異なる発音の「パン」と暫定的に認識し、その次に正しく「炭酸」と認識していることがわかる。`reply_to` というのは、そのタイミングでプログラムが Chrome ブラウザに対し、発話指示をしたことを示している。つまり図4ならば、「炭酸水」という単語を認識した1163ミリ秒の1ミリ秒後には、そのセリフに対する応答をブラウザに指示している。

#### 4. 実験結果と分析



図5 実験風景

実験はすべて1台の、Windows PC上 (NEC製 LAVIE, Intel® Core™ i5-6200U CPU @ 2.30GHz に Windows10 を搭載。オーディオは Realtek High Definition Driver を利用)で行った。これにより、同じパラメータを設定した時の、各被験者に対する遅延や音量は、共通に保たれている。

図5が実験風景である。実験は、20代後半から50代までの13名(男:7名, 女:6名)に対して行った。実験の際は、実験者1名と、被験者1名, 補助者1名が同室に入り、実験者が、被験者に対して全体説明のあと、SET A から順に試して記録をするように指示をした。各被験者の実験時間はそれぞれ約20分であった。

以下で、各セット内での各パラメータ設定の種類と、それぞれの設定に対して被験者が、1(心地よい)から4(不快)の間でどのような評価をつけたかを示す。

##### 4.1 実験の設定:

- インターバル: 400ms を単位として3段階に設定。0.0秒が最速のレスポンスで、1.2秒, 2.4秒に設定。一番遅い発話は、一番早い場合と比較して2.4秒遅れる。
- スピード: 1, 2, 4の3段階。1が最も遅く、4が最速
- ピッチ: 0, 0.9, 1.8の3段階。0が最も低く1.8が最も高い声

SET A から SET C は、発話速度を3段階に設定し、それぞれの場合において、インターバルを0.0秒, 1.2秒, 2.4秒の3段階に変化させて相違による印象の相違を確認する。以下、インターバルは秒単位で記載している。それぞれのパターンにおけるパラメータ設定をまとめると、次のようになる。

- SET A: 発話スピードはもっとも遅い。ピッチは中程度。  
 SET B: SET A と比較して発話スピードがやや速い  
 SET C: SET B よりも、さらに発話スピードが速い。

	パターン	インターバル	スピード	ピッチ
SET A	1	0.0	1	0.9
	2	1.2	1	0.9
	3	2.4	1	0.9
SET B	4	0.0	2	0.9
	5	1.2	2	0.9
	6	2.4	2	0.9
SET C	7	0.0	4	0.9
	8	1.2	4	0.9
	9	2.4	4	0.9

ここまでの設定で、被験者は SET A から SET C それぞれの中の組み合わせに対して評価を行うことになるが、ここで追加として、「SET A ~ SET C それぞれの中でもっとも心地よいと感じたインターバルのもの同士を比較し、その中でもっとも心地よいものはどれか？」という質問も行う。

次に、SET D と SET E では、発話スピードは中程度に固定し、ピッチを変えて、影響を調べる。

- SET D: ピッチは低めで、インターバルとスピードは B と同様。ただし、選択肢の並び順は SET C とは逆。  
 SET E: SET D と同様だが、ピッチが高め。



	パターン	インターバル	スピード	ピッチ
SET D	10	2.4	2	0
	11	1.2	2	0
	12	0.0	2	0
SET E	13	2.4	2	1.8
	14	1.2	2	1.8
	15	0.0	2	1.8

## 4.2 実験結果

以上の結果、被験者一人に対して合計 30 個のパラメータ設定の評価結果が得られる。それを 13 名の被験者に対して集計し、いくつかの角度から分析した結果を下記に示す。なお 1 (心地よい) から 4 (不快) までの 4 つの選択肢があったが、回答の分布は図 6 のようになった。

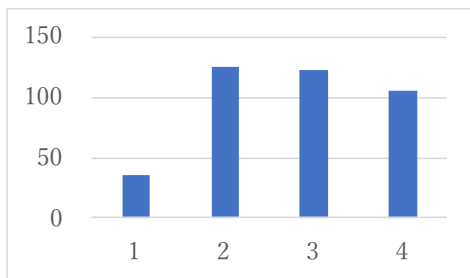


図 6 被験者の評価の分布

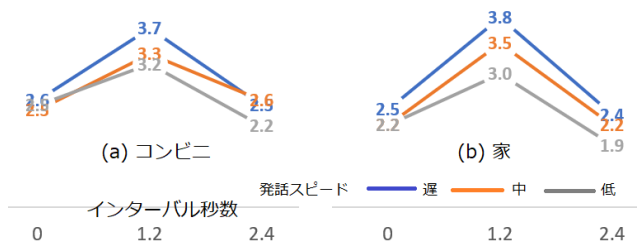


図 7 発話スピードと、心地よいインターバル  
(ピッチは 0.9 で、縦軸は心地よさの評価値)

図 7 は、13 人の被験者が、SET A から SET C に対して付けた心地よさの平均値を示している。いずれもピッチは 0.9 に固定されており、インターバルを 0.0 から 2.4 まで変化させたときに、心地よさの評価が、発話スピードによってどのように異なるかを示している。もっともスコアが高いのは、インターバルを 1.2 に設定した時である。どちらの場合も、最も発話スピードが遅い場合のスコアが最も高くなっている。発話スピードが最も速い発声は全般にスコアが低く、特に、発話スピードが速くインターバルが長い場合、つまり「レスポンスが遅い割に、話すスピードが速い」という場合は非常にスコアが低い。全体として、発話スピードよりはインターバルの方が、心地よさに対する影響は大きいように見える。

同じデータを、シーンとインターバルとの関係で示す。

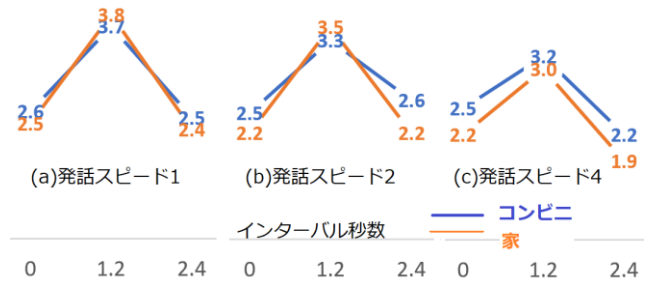


図 8 シーンと、心地よいインターバル  
(ピッチは 0.9 で、縦軸は心地よさの評価値)

図 8 の 3 つのグラフを見ると、「コンビニ」のシーンの方が、心地よさがインターバルの変化に対して敏感でない傾向にある。例えば、グラフ (a) では、「家」の平均スコアは、インターバルが 1.2 秒に設定されたときの「コンビニ」の平均スコアよりも高いが、他のインターバルでは、「家」のスコアが「コンビニ」よりも低い。発話スピードが速い時も同じ傾向である。我々は、これは、シーンおよび発話内容によって引き起こされると考えている。コンビニでは、情報を得るためだけに会話を行うが、「家」での会話は感情に関わるものだからである。

上記では、話すスピードによってインターバルの心地よさが変わってくることを示したが、さらにそれが、ピッチによっても変わってくるのではないかと仮説のもとに、ピッチの影響を調べた。ピッチは音の高さに相当するので、一般的に、ピッチが 0 の時には男性の声のように聞こえ、ピッチが 1.8 の時は女性の声のように聞こえる。

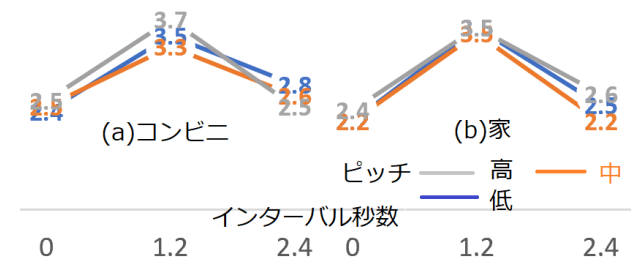


図 9 ピッチと、心地よいインターバル  
(スピードは 3 で、縦軸は心地よさの評価値)

SET D と SET E において、いずれもスピードは中間の 2 に固定しているので、同じくスピードを 2 としている SET B と合わせて比較をしたものが図 9 である。3 つの線はそれぞれ、ピッチが 0 (低い声)、ピッチが 0.9 (中間)、ピッチが 1.8 (高い声) の場合に対応している。このグラフによれば、ピッチは心地よさに対して、インターバルほどは影響を与えていないが、全体として、コンビニのシーンでは高いピッチの方が低いピッチよりも心地よさが高めとなっている。以上から、次のようなことが言える。

- 一般に、発話レスポンスの適切なインターバルというのは存在し、それよりも早めのレスポンスでも遅めのレスポンスでも、心地よさは低下する。
- 発話スピードは、心地よさに対してインターバルほどではないが、ある程度影響する。たとえば、発話スピードが速すぎると心地よさは低下する。ピッチも影響するが、その影響の度合いは、シーン（シチュエーション）に依存する傾向がある。これは、シーンによって適したピッチというのがあることを示唆する。
- インターバルの相違による影響の度合いは、シーンや会話の内容にも依存する。個人的、あるいは感情に関わる会話の方が、事務的な会話よりもインターバルの相違に対して敏感な傾向がある。

ここまでは発話スピードやピッチと、心地よいインターバルとの関係を見てきたが、「それでは、心地よい発話スピードはどれくらいなのか？」という疑問がある。筆者等は、それには各個人の発話スピードに関係していると考え、各個人が、実験における最初の発話である「炭酸水はありますか？」という発話にかかった時間を測定し、その数値と、その人が心地よいと感じるスピーカーの発話速度とを比較したが、結果として、各ユーザの発話速度と心地よい発話スピードとの間には、明確な相関関係が見出されなかった。つまり、「話すスピードが速い人が、システムの発話は遅い方を好む」というケースもあり、逆も成り立つ。

なお、上記のように、各ユーザ自身の好むスピーチ速度は、年齢層（20代、30代、40代、50代）および性別との間には、ある程度の関連性が見られた（母数が不十分で、統計的な検証はできていない）。男性と高齢の被験者はより速いスピーチを好む傾向があったのに対し、若年者と女性の被験者はより遅いスピーチを好む傾向が見られた。

## 5. 考察

一般に VUI システムにおいて、仮想的に設定されているキャラクターとしての一貫性を保つならば、発話によってピッチを大きく上下させるわけにはいかない。一方、内容によって、話すスピードを変えたり、インターバルを変えたりすることは人間の会話でも存在し、違和感は少ない。実験結果によれば、ピッチを固定したとして、発話スピードを想定する利用者層によって変えたり、個々の利用者の属性がわかる場合であれば、パーソナライズすることも考えられる。発話のためのインターバルは、その値については、さらに細かい検討が必要なものの、本論文で述べた、発話終了から 1.2 秒というのは一つの目安になるだろう。

参考になる事例として、ロボットの分野で石黒が行った実験がある。石黒はロボット演劇において、劇作家の平田オリザと共同で演出をし、その中で平田が「0.2 秒遅くし

て」などの指示をするだけで、リアリティが大きく向上したと報告している[4]。これは演劇で、あらかじめ定まったセリフを言う際の制御であるが、VUI の制御に応用する場合には、その時々で柔軟なセリフに対し、動的に発話タイミングの制御をする必要がある。

## 6. おわりに

Voice UI の心地よさを向上する方法として、これまで扱われていなかった時間要素の制御について述べた。Voice UI は今後、大きく普及して日常的に使われるようになることが想定されており、それらのシステムのユーザエクスペリエンス向上に大きく貢献すると考えている。

**謝辞** 被験者として実験に協力いただいた（株）サイバードの皆さまに感謝いたします。

## 参考文献

- [1] BBC article, June 9, 2014, Computer AI passes Turing test in 'world first', <https://www.bbc.com/news/technology-27762088>
- [2] Canalys report, May 23, 2018, Google beats Amazon to first place in smart speaker market, <https://www.canalys.com/newsroom/google-beats-amazon-to-first-place-in-smart-speaker-market>
- [3] CNET article, August 2, 2018, Apple's HomePod so far an also-ran with 6 percent of smart speaker market, <https://www.cnet.com/news/apples-homepod-so-far-an-also-ran-with-6-percent-of-smart-speaker-market/>
- [4] 石黒浩, 平田オリザ: ロボット演劇, 日本ロボット学会誌, Vol.29, No.1, pp.35-38, 2011
- [5] 神場知成, 2018, チャットボットにおける時間要素の設計について, <http://www.interaction-ipsj.org/proceedings/2018/data/pdf/1P01.pdf>
- [6] MDN web docs mozilla, "Using the Web Speech API", [https://developer.mozilla.org/en-US/docs/Web/API/Web\\_Speech\\_API/Using\\_the\\_Web\\_Speech\\_API](https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API/Using_the_Web_Speech_API)
- [7] M. Porcheron et al., 2018 Voice Interfaces in Everyday Life, CHI 2018, Paper No. 640, <https://doi.org/10.1145/3173574.3174214>
- [8] S. Reeves et al., 2018 Voice-based Conversational UX Studies and Design, CHI EA '18, Paper No. W38, <https://doi.org/10.1145/3170427.3170619>
- [9] J.Y. Tsai et al., 2018, Alexa, play some music: Categorization of Alexa Command, <https://voiceux.files.wordpress.com/2018/03/tsai.pdf>
- [10] A.M. Turing, 1950, Computing Machinery and Intelligence, Mind 49, 433-460.
- [11] Web Speech API Specification, <https://dvcs.w3.org/hg/speech-api/raw-file/9a0075d25326/speechapi.html>
- [12] T. Winograd, 1972, Understanding Natural Language, Cognitive Psychology 3(1), Academic Press 1972, 1-191, ISBN 0127597506, [https://doi.org/10.1016/0010-0285\(72\)90002-3](https://doi.org/10.1016/0010-0285(72)90002-3)
- [13] Workshop at the CHI 2018 conference, held at the Palais des Congrès de Montreal, Voice-based Conversational UX Studies and Design., <https://voiceux.wordpress.com/>
- [14] Z. Yin et al., 2017, DeepProbe: Information Directed Sequence Understanding and Chatbot Design via Recurrent Neural Networks, KDD '17, 2131-2139, <https://doi.org/10.1145/3097983.3098148>